

SagE-QG: Skill-Aligned and Evidence-Grounded Educational Question Generation via Counterfactual Evidence Reinforcement Learning

Yuqiu Deng¹, Chaoli Zhang^{2*†}, Chenyu Liu³, Xinliang Zhou³, Xingjun Wei⁴

Stephen Wang⁵, Yanfeng Wang⁶, Haoyang Li¹, Qingsong Wen^{7*}

¹Squirrel Ai Learning, China; ²Zhejiang Normal University

³College of Computing and Data Science, Nanyang Technological University;

⁴Telfer School of Management, University of Ottawa

⁵ABEL Intelligence, Inc.; ⁶School of Artificial Intelligence, Shanghai Jiao Tong University

⁷Squirrel Ai Learning, USA

genuineyukeo@gmail.com; chaolizcl@zjnu.edu.cn; qingsongedu@gmail.com

Abstract

Automatic Question Generation (QG) has become increasingly important for reading-comprehension practice and assessment. For educational use, generated questions should diagnose a target comprehension skill and rely on the passage evidence an expert would use to write the question. Existing answer-aware and skill-conditioned QG methods improve controllability, but often leave these skill-to-question and evidence-to-question dependencies implicit. We present **SagE-QG**, a **Skill-aligned and Evidence-Grounding Framework for Question Generation**. **SagE-QG** exposes evidence-grounded trajectories, optimizes multi-aspect rewards for answerability, skill alignment, and human similarity, and applies Counterfactual Evidence Intervention in reinforcement learning. The intervention removes selected evidence and rewards the Interventional Grounding Gap between full-passage and masked-evidence generations. Experiments on three reading-comprehension QG datasets show that **SagE-QG** improves evidence sensitivity and achieves notable gains on most metrics.

1 Introduction

Automatic question generation (QG) can help educators create scalable reading-comprehension practice, formative assessment, and personalized learning materials (Gonzalez et al., 2023; Uto et al., 2023; Wang et al., 2025). As LLMs enter educational workflows, QG shifts from producing fluent questions to supporting skill diagnosis, feedback, and assessment. Reading-comprehension pedagogy commonly organizes such skills into fine-grained categories, including literal understanding, reorganization, inference, evaluation, and appreciation under Barrett’s taxonomy (Clymer, 1968).

*Corresponding authors.

[†]This work was supported in part by the NSFC under Grant No. 62502456.

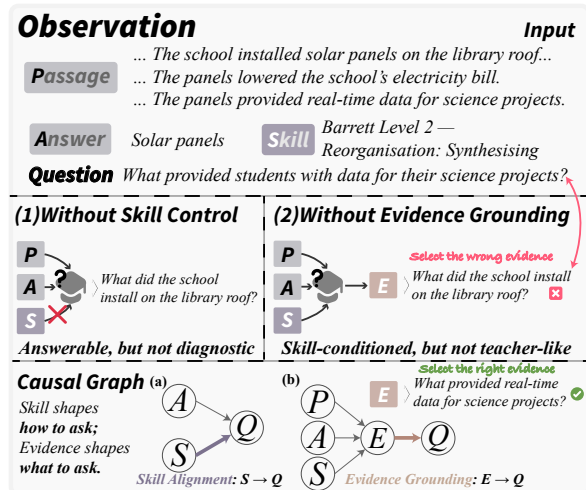


Figure 1: **Motivating observations of skill-aligned evidence-grounded QG.** Skill control determines how to ask, while evidence grounding determines what to ask; removing either factor leads to questions that are answerable but non-diagnostic, or skill-conditioned but not human-like.

This educational structure changes the target of QG from producing a plausible question to producing a question that tests the intended skill under a given passage and answer.

Prior QG work provides useful but incomplete controls. Answer-aware systems condition on passages and intended answers (Du et al., 2017; Yuan et al., 2017), sometimes with QA-oriented reinforcement rewards (Yuan et al., 2017). Educational QG adds skill cues through targeted inferential generation (Ghanem et al., 2022), SkillQG-style comprehension-type conditioning (Wang et al., 2023), and few-shot LLM prompting (Poon et al., 2024). Recent work also improves QG through diversity-seeking reinforcement learning (Ren et al., 2025) or test-time generation-and-filtering without updating the generator (Li and Zhang, 2025). Together, these methods push QG toward more controllable educational generation,

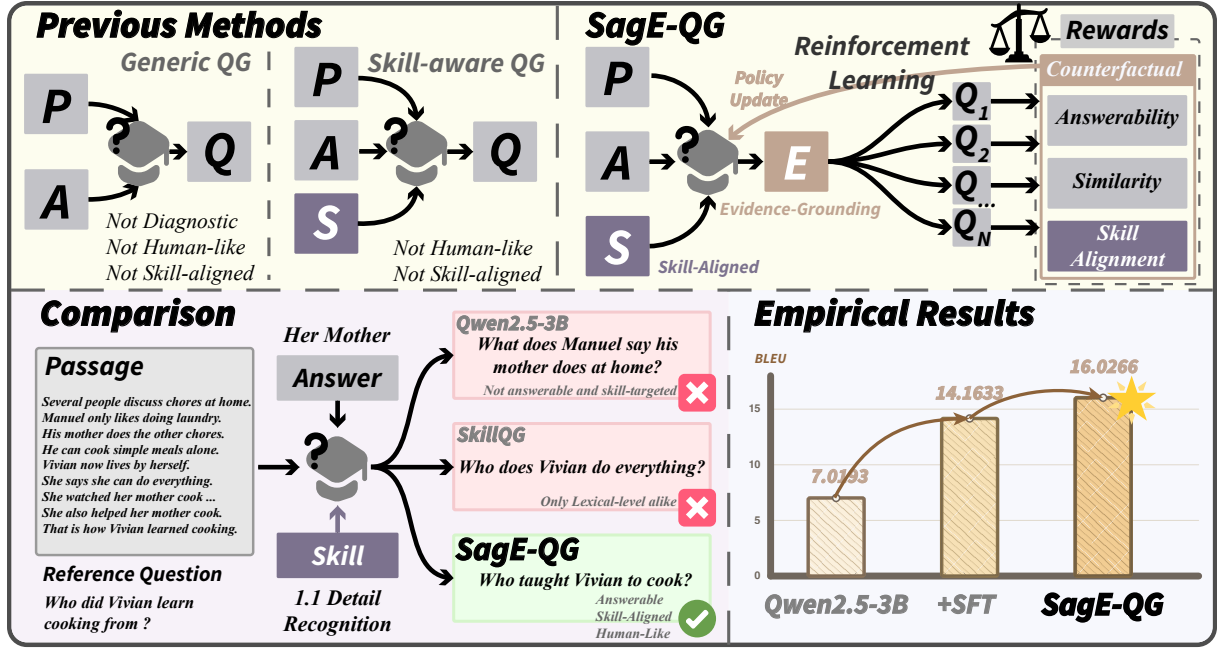


Figure 2: **Overview of Sage-QG.** Compared with prior generic or skill-aware QG methods, **SageE-QG** aligns generated questions with both supporting evidence and target skills through counterfactual reinforcement learning. The lower panels illustrate a representative comparison case and the progressive empirical gains achieved by our multi-stage training pipeline.

yet skill is at best treated as an input signal.

Figure 1 illustrates two common failures in educational QG. Without skill control, an answerable question may still be weakly diagnostic because the assessed skill is unclear. Even with skill conditioning, a model may target the wrong skill or rely on evidence that a human expert would not naturally select. These failures stem from prior work treating skill only as an input condition and evidence grounding as implicit context or a post-hoc rationale (Du and Cardie, 2017; Wang et al., 2023), rather than directly modeling the desired dependencies: $S \rightarrow Q$, where the **target skill** S controls what the question assesses, and $E \rightarrow Q$, where **selected evidence** E controls what the question asks. Figure 2 previews this shift from hidden-evidence generation to selected-evidence generation. This motivates our research question:

“How can question generation align with intended skills while remaining grounded in specified evidence?”

We propose **SageE-QG**, a **Skill-aligned** and **Evidence-Grounding Framework for Question Generation**. Given a passage P , an answer A , and a Barrett-style target skill S (Clymer, 1968), **SageE-QG** generates a structured trajectory: evidence scope, selected evidence, ques-

tion design, and final question. This trajectory makes selected evidence an explicit intermediate step, $(P, A, S) \rightarrow E \rightarrow Q$. Training proceeds in three stages. Skill-Aware Factual Grounding SFT teaches the model to produce skill-conditioned, evidence-grounded trajectories; multi-aspect alignment reward modeling evaluates answerability, target-skill consistency, and human similarity; and Counterfactual Evidence Reinforcement Learning (CE-RL) strengthens whether the generated question actually depends on the selected evidence. Specifically, CE-RL removes the selected evidence from the passage and uses the Interventional Grounding Gap between full and masked passage generations as a reinforcement signal. Experiments across three datasets show that **SageE-QG** improves evidence sensitivity and achieves notable gains on most metrics.

Our contributions are as follows:

- We formulate selected evidence as an explicit intermediate decision for expert-like question generation, making question-worthy evidence part of the generation target.
- We develop an LLM-based construction pipeline for Barrett-style skill labels and teacher-generated trajectories, yielding annotated splits for three reading-comprehension QG datasets.
- We propose **SageE-QG** with Counterfactual Evi-

dence Reinforcement Learning, using the Interventional Grounding Gap from evidence removal as a reinforcement signal to strengthen evidence-grounded, skill-aligned question generation.

2 Preliminaries

Task Definition. We study skill-aligned question generation for reading comprehension: given a passage P , an intended answer A , and a target comprehension skill S , the model is asked to generate a question whose answer is A and whose reasoning demand matches S . Each instance also includes a reference question Q^* . The target skill S is drawn from a Barrett-style closed taxonomy (Clymer, 1968), whose categories capture reading skills such as literal retrieval, reorganization, inference, evaluation, and appreciation. To obtain skill annotations, a teacher model assigns one taxonomy label to each reference question Q^* based on its passage and intended answer. We treat the assigned label as the target skill S for training and evaluation; details of the annotation protocol are described in the experimental setup.

For example, in a FairytaleQA passage about two neighbors, the intended answer A is “mow together and then divide the hay,” while the selected evidence E is the sentence stating this shared task. The generated question Q , “What task were the neighbors supposed to do together?”, targets A while using E to determine its passage focus.

Dependency Path View. Conventional answer-aware QG mainly relies on $(P, A) \rightarrow Q$. Skill-conditioned QG introduces S as an additional input, yet the generated question may still be dominated by passage-answer cues and fail to reflect the target skill. We further introduce E to denote the model-predicted evidence selection from P , represented as sentence indices, which captures the question-worthy content that should guide generation. The target dependency structure is summarized as

$$\begin{aligned} (P, A) &\rightarrow Q, \\ S &\rightarrow Q, \quad E \rightarrow Q, \quad E \subseteq P. \end{aligned} \quad (1)$$

This path view preserves answer-aware generation while making two educational dependencies explicit: S should control the reading skill, and E should control the content focus of the question.

Core Terminology. In the trajectory generated by **SagE-QG**, the **evidence scope** records the broader region considered, the **selected evidence** E records

the evidence used for masking and comparison, and the **question design** records how that evidence is turned into the **final question** Q under the target skill. Skill alignment is the property that the generated question tests the intended reading skill specified by S . Evidence Grounding is the property that the selected evidence supports the answer-question relation and affects the final question. The **Interventional Grounding Gap** is the score difference between full-passage generation and masked-evidence generation after removing the selected evidence.

3 Method

3.1 Overview

SagE-QG trains a large language model to generate questions through a skill-aware evidence process. Given (P, A, S) , **SagE-QG** produces a structured trajectory containing an evidence scope, selected evidence, a question design, and a final question. Figure 3 shows the three-stage pipeline: Skill-Aware Factual Grounding SFT (SFG-SFT) teaches the normal generation path $P, A, S \rightarrow E \rightarrow Q$; Multi-Aspect Alignment Reward Modeling (MA-RM) encourages sufficient answer support, skill consistency, and human similarity; and Counterfactual Evidence Reinforcement Learning (CE-RL) removes selected evidence and measures how much the question quality changes.

3.2 Skill-Aware Factual Grounding SFT

Skill-Aware Factual Grounding SFT (SFG-SFT) teaches the policy to serialize question generation as a skill-conditioned response trajectory before reinforcement learning. For each training instance, the annotation pipeline attaches an evidence scope, selected evidence, and a question-design rationale to the reference question target under the target skill. We denote the resulting target response as

$$Y^* = \langle C^*, E^*, D^*, Q^* \rangle, \quad (2)$$

where the tuple denotes four serialized fields in one response: C^* is the evidence scope, E^* is the selected evidence, D^* is the skill-aware question design, and Q^* is the final question. Given $X = (P, A, S)$, we optimize the standard cross-entropy objective

$$\mathcal{L}_{\text{SFT}} = - \sum_{t=1}^{|Y^*|} \log \pi_{\theta}(y_t^* \mid X, y_{<t}^*). \quad (3)$$

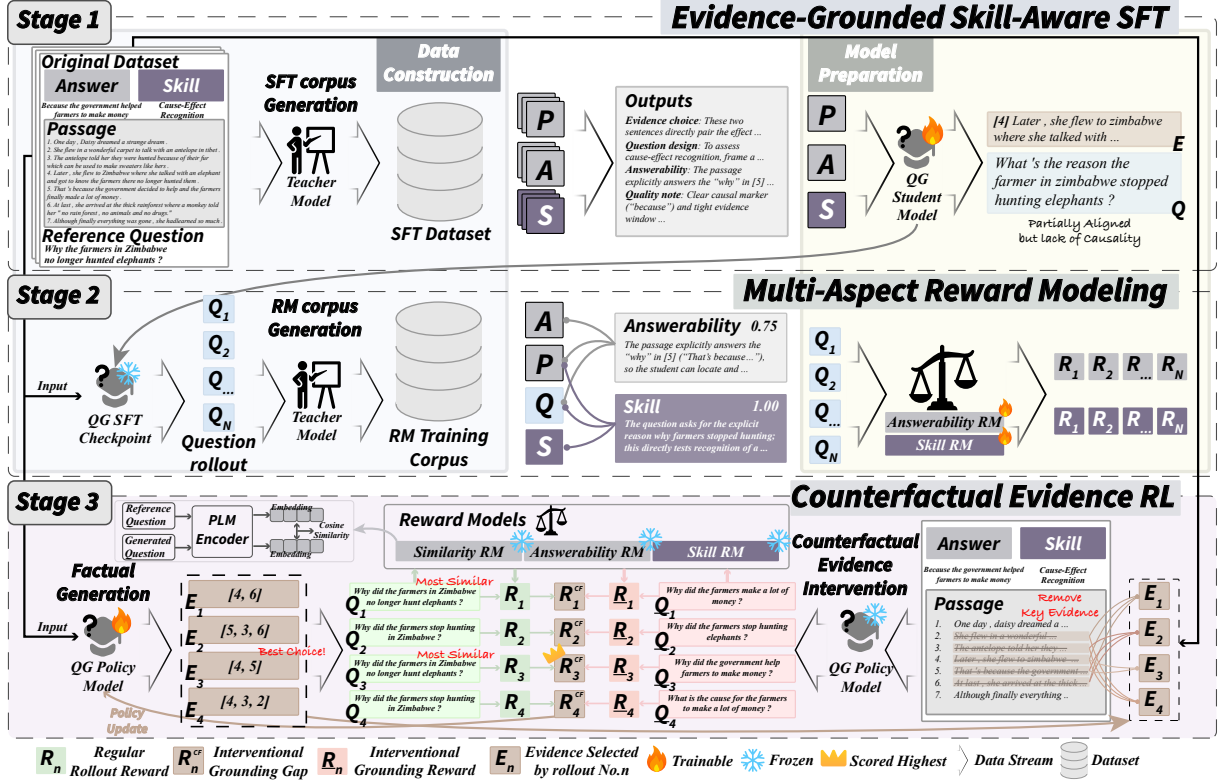


Figure 3: **Three-stage training pipeline of Sage-QG.** Sage-QG first uses Skill-Aware Factual Grounding SFT (SFG-SFT) to learn skill-conditioned evidence trajectories, then trains Multi-Aspect Alignment Reward Models (MA-RM) for answerability, skill alignment, and human similarity. Finally, Counterfactual Evidence Reinforcement Learning (CE-RL) removes the selected evidence and optimizes the policy with the Interventional Grounding Gap between full-passage and masked-evidence generations, encouraging questions that are both skill-aligned and evidence-grounded.

Since E^* and D^* are part of the target sequence, the model learns to expose which evidence is sufficient for the answer-question relation and how that evidence should be used under the target skill before generating the final question.

This SFT pass also creates the data source for later reward modeling. We roll out the SFT policy on training prompts to obtain candidate trajectories, then use teacher-model judgments and reference questions to construct multi-aspect supervision for the reward models.

3.3 Multi-Aspect Alignment Reward Modeling

Multi-Aspect Alignment Reward Modeling (MA-RM) assigns scalar scores to generated questions. The answerability and skill reward models use the same pretrained backbone as the SFT policy and are trained on SFT rollouts paired with teacher-model labels under fixed rubrics. A fixed pretrained encoder supplies the reference-similarity compo-

nent. We use three weighted reward components:

$$R_{\text{core}}(Q, P, A, S) = \lambda_{\text{ans}} R_{\text{ans}}(Q, P, A) + \lambda_{\text{skill}} R_{\text{skill}}(Q, S) + \lambda_{\text{sim}} R_{\text{sim}}(Q, Q^*). \quad (4)$$

Here, R_{ans} measures whether the question is answerable from the passage with the intended answer, R_{skill} measures closed-skill consistency, and R_{sim} measures human similarity through reference-question similarity. The coefficients λ_{ans} , λ_{skill} , and λ_{sim} control the relative strength of each reward; we use unit weights unless otherwise specified.

The reward components mainly score full-passage question quality. Answerability keeps the question tied to the passage-answer pair. Skill scoring checks whether answering the generated question naturally requires the target reading skill. Human similarity discourages unnatural wording and question designs that diverge from reference questions. CE-RL then complements these quality-oriented rewards by measuring the Interventional

Grounding Gap after selected evidence is removed.

3.4 Counterfactual Evidence Reinforcement Learning

Counterfactual Evidence Reinforcement Learning (CE-RL) optimizes whether the selected evidence affects the generated question with Counterfactual Evidence Intervention. For an input $X = (P, A, S)$, the policy first samples a group of G full-passage trajectories

$$\begin{aligned} Y_i^+ &\sim \pi_\theta(\cdot \mid P, A, S), \\ Y_i^+ &= \langle C_i, E_i, D_i, Q_i^+ \rangle, \quad i = 1, \dots, G. \end{aligned} \quad (5)$$

For each factual trajectory, we remove its selected evidence and construct an intervened passage

$$P_i^{-E} = \text{mask}(P, E_i), \quad (6)$$

then sample a paired intervened trajectory

$$\begin{aligned} Y_i^{-E} &\sim \pi_\theta(\cdot \mid P_i^{-E}, A, S), \\ Y_i^{-E} &= \langle C_i^{-E}, E_i^{-E}, D_i^{-E}, Q_i^{-E} \rangle. \end{aligned} \quad (7)$$

The paired score difference defines the Interventional Grounding Gap:

$$\begin{aligned} \Delta_{\text{CF}}^{(i)} &= R_{\text{core}}(Q_i^+; P, A, S) \\ &\quad - R_{\text{core}}(Q_i^{-E}; P_i^{-E}, A, S). \end{aligned} \quad (8)$$

A positive gap means that removing the selected evidence reduces generation quality under the same answer and skill constraints. A negative gap penalizes full-passage trajectories whose selected evidence does not support the generated question.

The training reward for each full-passage trajectory combines the Interventional Grounding Gap with a fixed format reward:

$$R_{\text{train}}^{(i)} = \lambda_{\text{CF}} \Delta_{\text{CF}}^{(i)} + R_{\text{fmt}}(Y_i^+), \quad (9)$$

where R_{fmt} preserves the parseable trajectory format and is added with fixed unit weight. Thus CE-RL introduces only one new reward coefficient, λ_{CF} , which controls the strength of the counterfactual signal.

We optimize the policy with GRPO (Shao et al., 2024). Full-passage trajectories form the optimization group, while masked-evidence trajectories only provide paired scores for $\Delta_{\text{CF}}^{(i)}$. Within each group, GRPO converts training rewards into relative advantages:

$$\hat{A}_i = \frac{R_{\text{train}}^{(i)} - \frac{1}{G} \sum_{j=1}^G R_{\text{train}}^{(j)}}{\text{std}_j(R_{\text{train}}^{(j)}) + \epsilon}. \quad (10)$$

Using these advantages, the policy is updated with a clipped objective

$$\tilde{\rho}_i = \text{clip}(\rho_i, 1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}}), \quad (11)$$

$$\begin{aligned} \mathcal{L}_{\text{GRPO}} &= -\frac{1}{G} \sum_{i=1}^G \min(\rho_i \hat{A}_i, \tilde{\rho}_i \hat{A}_i) \\ &\quad + \beta D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}), \end{aligned} \quad (12)$$

where $\rho_i = \pi_\theta(Y_i^+ \mid X) / \pi_{\theta_{\text{old}}}(Y_i^+ \mid X)$ is written at the sequence level for compactness. Masked-evidence trajectories are excluded from the policy group, so CE-RL changes the full-passage policy only through the Interventional Grounding Gap it assigns to each full-passage rollout.

CE-RL therefore separates question quality from evidence dependence. R_{core} scores whether a generated question satisfies the answer, skill, and similarity criteria, whereas the paired difference measures how those criteria change after the selected evidence is removed. The update rewards full-passage trajectories whose selected evidence is consequential, while the masked generation serves only as the comparison condition.

4 Experiments

4.1 Experimental Setup

Following the protocol in Figure 3, every system receives the same passage, intended answer, and target skill. The main text reports reference-based metrics and educational metrics, with dataset details in Appendix C, training hyperparameters and resources in Table 7, additional experiments in Appendix D, and human-alignment validation in Appendix D.7.

Datasets. We evaluate skill-aligned reading-comprehension question generation on EQG-RACE (Jia et al., 2021), DREAM (Sun et al., 2019), and FairytaleQA (Xu et al., 2022), covering exam-style passages, dialogues, and narratives. Each system is given a passage, a target answer, and a target reading skill, where the skill labels are Barrett-style annotations produced by our automatic annotation pipeline using GPT-5 (OpenAI, 2025). The source datasets do not contain gold evidence labels. Therefore, for **SagE-QG** training, the teacher pipeline constructs pseudo-evidence traces, and at inference time the model exposes its selected evidence as an intermediate decision.

These supervision sources serve different roles across training stages. The reference question supplies the SFT question target and the similarity

Table 1: Main results on reading-comprehension QG datasets under the 3B backbone setting. Educational metrics are judged by GPT-5 on a 0–1 scale. Arrows indicate metric direction. Bold marks the best result and underline marks the second best among filled entries when at least two systems have reported values.

Dataset	System	Educational Metrics			Reference-Based Metrics			
		Skill $\uparrow$	Answer. $\uparrow$	Human Sim. $\uparrow$	BLEU $\uparrow$	ROUGE-1 $\uparrow$	ROUGE-L $\uparrow$	BERTScore $\uparrow$
EQG-RACE	Qwen2.5-3B Base	<u>0.7606</u>	<u>0.6745</u>	0.5771	7.0193	31.3835	28.4679	0.7133
	SkillQG (3B)	0.6832	0.6121	0.5164	16.2084	<u>42.2016</u>	<u>40.3374</u>	<u>0.7385</u>
	SagE-QG	0.7644	0.6768	<u>0.5611</u>	<u>16.0266</u>	42.9198	41.0802	0.7542
DREAM	Qwen2.5-3B Base	0.6746	0.6884	0.6118	10.2120	34.6070	33.0555	0.6620
	SkillQG (3B)	<u>0.7244</u>	<u>0.7151</u>	0.6396	24.9486	50.6748	49.8644	<u>0.7669</u>
	SagE-QG	0.7775	0.7863	0.6850	29.7274	56.2337	55.3622	0.7996
FairytaleQA	Qwen2.5-3B Base	<u>0.7696</u>	<u>0.6714</u>	<u>0.6249</u>	13.7364	40.7823	38.7207	0.7383
	SkillQG (3B)	0.6897	0.6532	0.5687	<u>23.9072</u>	<u>49.7782</u>	<u>48.1980</u>	<u>0.7622</u>
	SagE-QG	0.7814	0.7252	0.6587	25.5100	50.5300	49.0000	0.7639

reward, while the teacher-generated trace supplies the evidence scope, selected evidence, and question design. At inference time, the model receives only (P, A, S) and generates the complete trajectory; neither the reference question nor teacher evidence is required.

Baselines. The main comparison uses Qwen2.5-3B (Qwen Team, 2024) as the shared backbone scale for **SagE-QG** and open-source LLM baselines. We compare against the base instruction model and SkillQG (Wang et al., 2023). For SkillQG, we replace its model components with Qwen2.5-3B while keeping the original training pipeline and supervision format unchanged. All LLM baselines use the same generation prompt template and receive the same passage, intended answer, and target skill; thus, differences from **SagE-QG** are not caused by giving baselines less input information. We report Qwen2.5-7B (Qwen Team, 2024) separately as a scaling analysis on EQG-RACE, and include a further comparison with GPT-5 (OpenAI, 2025) in Appendix D.5.

Metrics. We report BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2019) for reference similarity. Educational metrics use GPT-5 as a fixed LLM judge of target-skill diagnosis, answerability, and human similarity in question design on a 0–1 scale. The Human Sim. score evaluates the generated question’s wording and questioning angle against human-written questions, separate from lexical overlap with a reference. When reported, Joint is the sum of the three aspect scores. We report the human-alignment validation in Appendix D.7 and the prompt for LLM-judge in Appendix D.1. All reported automatic

metrics are obtained from a single run.

4.2 Main Results

Table 1 reports the main comparison under the shared 3B backbone setting, keeping the parameter scale fixed for a fair comparison. SkillQG uses its original three-stage inference procedure, while the base model and **SagE-QG** generate each question in a single pass. **SagE-QG** gives the strongest target-skill and answerability scores on all three datasets. It also gives the strongest human-similarity scores on DREAM and FairytaleQA, while the base model is slightly higher on EQG-RACE.

Across the reference-based metrics, **SagE-QG** gives the strongest overall pattern on DREAM and FairytaleQA and remains competitive on EQG-RACE. The FairytaleQA results follow this pattern: **SagE-QG** leads the reference-based comparison, with SkillQG still close on lexical overlap. Its higher educational scores indicate stronger target-skill diagnosis and more human-like question design under the LLM judge. These results support reporting reference similarity and educational judgment together, since they capture complementary aspects of question quality.

The base instruction model has weak reference-based scores because it has not learned the dataset-specific reference distribution. Still, it remains competitive on several educational metrics, likely because preference alignment gives the backbone a useful prior for fluent and answerable questions. Relative to this base model, **SagE-QG** improves all reference-based metrics and all target-skill and answerability scores. Relative to SkillQG, **SagE-QG** gives higher judge-based educational scores

Table 2: Ablation study over training components. **SagE-QG** is the full setting with counterfactual evidence reinforcement learning. Educational metrics are judged by GPT-5 on a 0–1 scale.

Dataset	Component	Educational Metrics			Reference-Based Metrics			
		Skill $\uparrow$	Answer. $\uparrow$	Human Sim. $\uparrow$	BLEU $\uparrow$	ROUGE-1 $\uparrow$	ROUGE-L $\uparrow$	BERTScore $\uparrow$
EQG-RACE	Base	0.7606	0.6745	0.5771	7.0193	31.3835	28.4679	0.7133
	+ SFT	0.7213	0.6409	0.5409	14.1633	40.6490	39.1367	0.7399
	+ GRPO	0.7195	0.6544	0.5463	13.1897	39.0914	37.2885	0.7330
	SagE-QG	0.7644	0.6768	0.5611	16.0266	42.9198	41.0802	0.7542
DREAM	Base	0.6746	0.6884	0.6118	10.2120	34.6070	33.0555	0.6620
	+ SFT	0.7708	0.7776	0.7036	29.6582	56.0731	55.2429	0.7990
	+ GRPO	0.7475	0.7638	0.6863	29.4338	56.0147	55.1542	0.7986
	SagE-QG	0.7775	0.7863	0.6850	29.7274	56.2337	55.3622	0.7996
FairytaleQA	Base	0.7696	0.6714	0.6249	13.7364	40.7823	38.7207	0.7383
	+ SFT	0.7438	0.7171	0.6374	24.9675	50.1275	48.5869	0.7263
	+ GRPO	0.7767	0.7160	0.6530	26.4503	48.0681	46.6353	0.7115
	SagE-QG	0.7814	0.7252	0.6587	25.5100	50.5300	49.0000	0.7639

across the three datasets.

4.3 Ablation Study

Table 2 reports the incremental effect of each training component. The +SFT row shows that supervised trajectory learning captures much of the dataset-specific QG pattern by imitating teacher-constructed evidence scopes, selected evidence, question designs, and final questions. This leads to strong reference-based metrics, but SFT still does not explicitly test whether the selected evidence is necessary for the generated question.

The +GRPO row is an RL control: it shares the backbone, data, SFT initialization, (P, A, S) input, and multi-aspect rewards with **SagE-QG**, while omitting the evidence intervention and Interventional Grounding Gap. It optimizes question-level quality on full-passage generations, but is less likely to strengthen the dependency between selected evidence and question generation. **SagE-QG** adds the counterfactual evidence comparison, making the selected evidence more likely to become consequential during optimization.

4.4 Experimental Analysis

We further conduct analyses to better understand the results in the main tables. We examine scaling behavior, evidence sensitivity under Counterfactual Rollout, and the relationship between reward components and external metrics.

4.4.1 Scaling Analysis

We analyze model scale by comparing Qwen2.5-3B and Qwen2.5-7B on EQG-RACE. Table 3 reports

matched rows for each training variant.

The scaling results show a stable pattern: moving from 3B to 7B improves every reported metric within each training variant. At 7B, **SagE-QG** obtains the best result on seven of eight metrics and outperforms the matched +SFT and +GRPO variants on every metric. These results support scaling the approach to larger models and broader training data.

4.4.2 Evidence Sensitivity Analysis

We use the Interventional Grounding Gap as a diagnostic of sensitivity to selected evidence; educational and reference-based metrics separately evaluate question quality.

Figure 4 compares how much each model’s reward components drop after its selected evidence is removed. For Qwen2.5-3B, we use a one-shot format prompt, because the zero-shot model does not reliably expose selected evidence in the re-

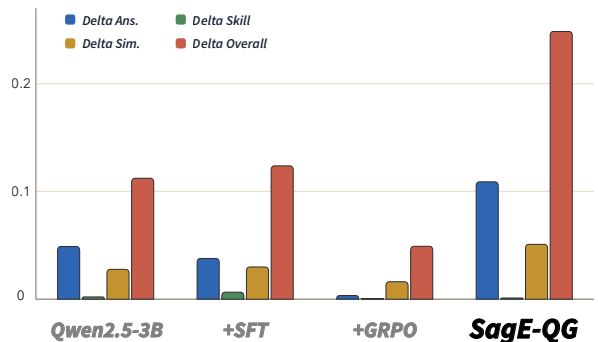


Figure 4: Interventional grounding gap breakdown after evidence removal; the Base row uses a compact one-shot format prompt.

Table 3: Scaling analysis on EQG-RACE by training variant. Each variant reports matched 3B and 7B rows. Skill, Answer., and Human Sim. are LLM judge scores on a 0–1 scale, and Joint is their sum. Bold marks the better score within each pair.

Variant	Parameters	Educational Metrics				Reference-Based Metrics			
		Skill ↑	Answer. ↑	Human Sim. ↑	Joint ↑	BLEU ↑	ROUGE-1 ↑	ROUGE-L ↑	BERTScore ↑
Base	3B	0.7606	0.6745	0.5771	2.0122	7.0193	31.3835	28.4679	0.7133
	7B	0.8004	0.7460	0.6288	2.1751	7.9782	31.7784	29.1406	0.7192
+ SFT	3B	0.7213	0.6409	0.5409	1.9031	14.1633	40.6490	39.1367	0.7399
	7B	0.7802	0.7021	0.6118	2.0940	17.5466	45.2903	43.5803	0.7691
+ GRPO	3B	0.7195	0.6544	0.5463	1.9203	13.1897	39.0914	37.2885	0.7330
	7B	0.7860	0.7116	0.6096	2.1073	17.8396	44.9582	43.2312	0.7659
SagE-QG	3B	0.7644	0.6768	0.5611	1.9724	16.0266	42.9198	41.0802	0.7542
	7B	0.8151	0.7442	0.6350	2.1942	18.2148	45.5638	43.8313	0.7702

Table 4: Reward–metric alignment analysis across saved EQG-RACE checkpoint evaluations.

Metric Pair	N	Pearson	Spearman
Overall Reward vs BLEU	218	0.4348	0.1728
Overall Reward vs ROUGE-1	218	0.6177	0.2558
Overall Reward vs ROUGE-L	218	0.3303	0.2439
Overall Reward vs BERTScore	218	0.7859	0.4540

quired output format. This formatted base probe already shows a moderate overall gap, mainly from answerability and similarity. The +SFT model gives a slightly larger overall gap, suggesting that supervised evidence trajectories teach evidence-conditioned generation patterns. +GRPO reduces this gap, indicating that full-passage reward optimization alone does not necessarily make selected evidence more consequential.

SagE-QG changes this pattern: its overall Interventional Grounding Gap reaches 0.2484, compared with 0.1237 for +SFT and 0.0491 for +GRPO. The largest component drop is answerability, suggesting that CE-RL makes selected evidence more consequential for whether the generated question remains answer-supported. Full numeric results are reported in Appendix D.6.

4.4.3 Reward–Metric Alignment Analysis

Table 4 reports how the overall reward aligns with reference-based metrics. The reward has positive correlation with all reported text-similarity metrics, with the strongest association on BERTScore. This indicates that the reward signal generally tracks external quality trends, especially semantic similarity.

These correlations are positive but uneven, which is expected because the overall reward combines answerability, skill alignment, and similarity rather than serving as a pure reference-overlap ob-

jective. The result supports using reward scores as a quality signal, while still reporting reference-based metrics and educational metrics separately.

5 Related Work

LLMs in Education. Recent work studies how LLMs support educational tasks such as tutoring, multimodal learning, agent-based teaching systems, and assessment design (Wang et al., 2025; Chu et al., 2025b; Ye et al., 2025; Chu et al., 2025a; Xu et al., 2025). A parallel line evaluates whether LLM-generated questions move beyond memorization toward deeper cognitive demand (Wang et al., 2026). Our work follows this direction but focuses on reading-comprehension QG, where question quality depends on both the target skill and the passage evidence used to ground the question.

Educational Question Generation. Within QG, neural systems are commonly formulated as passage- or passage-answer-conditioned generation (Du et al., 2017; Yuan et al., 2017), with reinforcement learning sometimes used to connect questions to downstream QA rewards (Yuan et al., 2017). Educational QG adds assessment-oriented signals, including inferential question generation (Ghanem et al., 2022), SkillQG-style comprehension-type conditioning (Wang et al., 2023), few-shot LLM prompting (Poon et al., 2024), diversity-seeking reinforcement learning (Ren et al., 2025), and test-time generation-and-filtering for difficulty consistency (Li and Zhang, 2025). SoftSkillQG controls reading skills through learned soft prompts, while Leite and Cardoso use textual attributes to control question types (Ohe and Fyshe, 2024; Leite and Cardoso, 2023). **SagE-QG** keeps the answer-aware setting, but makes both the target skill and

selected evidence explicit. CE-RL shares the use of paired conditions with contrastive learning (Khosla et al., 2020), while its signal is a sequence-level generation-quality difference rather than a representation-distance objective.

Evidence Focus and Evaluation. Prior work on reading-comprehension QG argues that passage-level systems should identify question-worthy sentences before generation (Du and Cardie, 2017). Evaluation studies further show that similarity to ground-truth question-answer pairs can bias automatic evaluation and that existing metrics may align poorly with human judgments (Moon et al., 2024; Fu et al., 2024). We connect these concerns through counterfactual evidence dependence: an assessment-quality question should change when its selected evidence is removed.

6 Conclusion

We present **SagE-QG**, a framework for skill-aligned and evidence-grounded reading-comprehension QG that introduces selected evidence as an explicit intermediate decision. By comparing full-passage generation with paired masked-evidence generation, CE-RL strengthens whether generated questions depend on the selected evidence. Experiments across three datasets show that **SagE-QG** improves evidence sensitivity and achieves notable gains on most metrics.

Limitations

SagE-QG is evaluated under a limited experimental scale. Although the experiments cover three reading-comprehension QG datasets, larger backbones, broader training data, and stronger training budgets are needed to test how far the gains persist. The evidence-intervention design is also intentionally simple: we remove the selected evidence spans with a fixed masking strategy and require non-empty evidence through the format reward. This keeps the method interpretable and easy to reproduce, but future work could study more compact evidence selection, alternative interventions, and controls that separate evidence dependence from answer removal.

Our evaluation relies on GPT-5 as an automatic judge for educational metrics. Appendix D.7 reports human-alignment validation, but judge bias may still affect which question styles receive higher scores. Future dataset construction should include

teacher- or domain-expert-curated skill labels so that training labels reflect the pedagogical interpretation used in deployment. Finally, **SagE-QG** is designed for reading-comprehension QG, where questions can be grounded in explicit textual evidence. It may be less directly applicable to tasks such as mathematical problem generation or open-ended assessment design, where quality depends more on abstract reasoning than on identifiable passage evidence.

Ethical Considerations

The experiments use public benchmark datasets and do not introduce new personally identifying student data. The public datasets were used under their existing research-use terms, and the curation instructions stated that derived annotations would be used for research and evaluation only. We use GPT-5 as an automatic judge for educational metrics, and Appendix D.7 reports a human-alignment validation for this judging protocol. AI assistants were used only for code debugging and paper polishing; all suggestions and changes were reviewed by the authors.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 62502456.

References

- Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jingheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S. Yu, and Qingsong Wen. 2025a. [LLM agents for education: Advances and applications](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13782–13810, Suzhou, China. Association for Computational Linguistics.
- Zhendong Chu, Jian Xie, Shen Wang, Zichao Wang, and Qingsong Wen. 2025b. [UniEDU: Toward unified and efficient large multimodal models for educational tasks](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1007–1016, Suzhou (China). Association for Computational Linguistics.
- Theodore Clymer. 1968. [What is “reading”? Some current concepts](#). *Teachers College Record: The Voice of Scholarship in Education*, 69:7 – 29.
- Xinya Du and Claire Cardie. 2017. [Identifying where to focus in reading comprehension for neural question generation](#). In *Proceedings of the 2017 Conference*

- on *Empirical Methods in Natural Language Processing*, pages 2067–2073, Copenhagen, Denmark. Association for Computational Linguistics.
- Xinya Du, Junru Shao, and Claire Cardie. 2017. [Learning to ask: Neural question generation for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.
- Weiping Fu, Bifan Wei, Jianxiang Hu, Zhongmin Cai, and Jun Liu. 2024. [QGEval: Benchmarking multi-dimensional evaluation for question generation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11783–11803, Miami, Florida, USA. Association for Computational Linguistics.
- Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer von der Ohe, and Alona Fyshe. 2022. [Question generation for reading comprehension assessment by modeling how and what to ask](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2131–2146, Dublin, Ireland. Association for Computational Linguistics.
- Hannah Gonzalez, Liam Dugan, Eleni Miltsakaki, Zhiqi Cui, Jiaxuan Ren, Bryan Li, Shriyash Upadhyay, Etan Ginsberg, and Chris Callison-Burch. 2023. [Enhancing human summaries for question-answer generation in education](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 108–118, Toronto, Canada. Association for Computational Linguistics.
- Xin Jia, Wenjie Zhou, Xu Sun, and Yunfang Wu. 2021. [EQG-RACE: Examination-Type Question Generation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13143–13151.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. 2020. [Supervised contrastive learning](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.
- Bernardo Leite and Henrique Lopes Cardoso. 2023. [Towards enriched controllability for educational question generation](#). In *Artificial Intelligence in Education: 24th International Conference, AIED 2023, Tokyo, Japan, July 3–7, 2023, Proceedings*, page 786–791, Berlin, Heidelberg. Springer-Verlag.
- Kunze Li and Yu Zhang. 2025. [CrossQG: Improving difficulty-controllable question generation through consistency enhancement](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2783–2798, Suzhou, China. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Hyeonseok Moon, Jaewook Lee, Sugyeong Eo, Chanjun Park, Jaehyung Seo, and Heuseok Lim. 2024. [Generative interpretation: Toward human-like evaluation for educational question-answer pair generation](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2185–2196, St. Julian’s, Malta. Association for Computational Linguistics.
- Spencer McIntosh von der Ohe and Alona Fyshe. 2024. [Softskillqg: Robust Skill-Targeted Reading Comprehension Question Generation Using Soft-prompts](#). *Proceedings of the Canadian Conference on Artificial Intelligence*. <https://caiac.pubpub.org/pub/n3kbtb8>.
- OpenAI. 2025. [GPT-5 System Card](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Yin Poon, John Sie Yuen Lee, Yu Yan Lam, Wing Lam Suen, Elsie Li Chen Ong, and Samuel Kai Wah Chu. 2024. [Few-shot question generation for reading comprehension](#). In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 21–27, Bangkok, Thailand. Association for Computational Linguistics.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Tianyu Ren, Hui Wang, and Karen Rafferty. 2025. [Enhancing question generation through diversity-seeking reinforcement learning with bilevel policy decomposition](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25083–25091.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. [DREAM: A challenge data set and models for dialogue-based reading comprehension](#). *Transactions of the Association for Computational Linguistics*, 7:217–231.
- Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2023. [Difficulty-controllable neural question generation for reading comprehension using item response theory](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*

- (BEA 2023), pages 119–129, Toronto, Canada. Association for Computational Linguistics.
- Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2025. [Large language models for education: A survey and outlook](#). *IEEE Signal Processing Magazine*, 42(6):51–63.
- Xiaolong Wang, Zhe Zhao, Song Lai, Chaoli Zhang, Zijie Geng, Yu Tong, Ye Wei, and Qingsong Wen. 2026. [From memorization to creation: Evaluating the cognitive depth of LLM-generated educational questions](#). In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2818–2829.
- Xiaoqiang Wang, Bang Liu, Siliang Tang, and Lingfei Wu. 2023. [SkillQG: Learning to generate question for reading comprehension assessment](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13833–13850, Toronto, Canada. Association for Computational Linguistics.
- Tianlong Xu, Yi-Fan Zhang, Zhendong Chu, and Qingsong Wen. 2025. [Multimodal AI teacher: Integrating edge computing and reasoning models for enhanced student error analysis](#). *AI Magazine*, 46(3):e70030.
- Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, Tran Bao Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. [Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.
- Jingheng Ye, Shen Wang, Deqing Zou, Yibo Yan, Kun Wang, Hai-Tao Zheng, Ruitong Liu, Zenglin Xu, Irwin King, Philip S. Yu, and Qingsong Wen. 2025. [Position: LLMs can be good tutors in English education](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17516–17535, Suzhou, China. Association for Computational Linguistics.
- Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordani, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. [Machine comprehension by text-to-text neural question generation](#). In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. [BERTScore: Evaluating Text Generation with BERT](#). *arXiv preprint*. Version Number: 3.

A Reproducibility Statement

We release the implementation, prompts, and evaluation scripts at <https://github.com/wythedee/SagE-QG>. Model checkpoints and processed datasets are available at <https://huggingface.co/wythedee/SagE-QG>. Dataset construction details, split statistics, evaluation protocols, and additional experimental results are documented in Appendix C and Appendix D. All experiments use the model scales and judging protocol described in Section 4.1.

B Artifact Availability and Licenses

We use public benchmark datasets and publicly available model checkpoints for research and evaluation purposes. The released checkpoints and datasets are hosted at <https://huggingface.co/wythedee/SagE-QG>. Users of our released code, prompts, preprocessing scripts, and processed annotations should follow the licenses and terms of the original datasets and model providers. The processed annotations and pseudo-evidence trajectories are released for research use only and should not be used as standalone educational assessments without human review.

C Dataset

The source datasets used in this paper provide reading passages, reference questions, and answer information. They do not include gold evidence labels. We assign Barrett-style skill labels with the GPT-based annotation pipeline described in the main text and construct teacher-generated pseudo-evidence trajectories for **SagE-QG** training. These trajectories support supervised trajectory learning and evidence-removal training; evaluation uses the normalized passage, answer, target skill, and question interface. Table 5 reports the option-clean annotated JSONL splits used for dataset-level reporting and evaluation. The resulting label distribution is naturally imbalanced, because easier reading skills occur more frequently than higher-order skills in the original data.

C.1 Dataset Details

EQG-RACE. EQG-RACE is an examination-style QG benchmark derived from RACE (Jia et al., 2021). It is built from English reading-comprehension examinations and emphasizes long passages, answer guidance, and questions that often require reasoning over multiple sentences.

DREAM. DREAM is a dialogue-based reading-comprehension dataset collected from English-as-a-foreign-language examinations (Sun et al., 2019). Its questions stress multi-turn dialogue understanding, implicit speaker intent, and reasoning beyond a single sentence.

FairytaleQA. FairytaleQA is a narrative-comprehension dataset built around story passages and educational questions (Xu et al., 2022). It tests whether generated questions remain meaningful when the source text is a coherent story instead of an exam passage or dialogue.

Together, these datasets cover examination-style, dialogue-based, and narrative QG settings, providing a broader basis for evaluating whether skill control and evidence grounding remain stable across different reading-comprehension scenarios.

C.2 Preprocessing

EQG-RACE converts the original multiple-choice RACE format into a free-form QG setting. Each RACE item contains a passage, a question, four answer options, and one correct option. Following the EQG-RACE construction, distractors are discarded and the correct option is retained as the target answer, yielding a ⟨passage, answer, question⟩ triple for answer-aware question generation (Jia et al., 2021). EQG-RACE further removes cloze-style and general-purpose questions, keeping specific questions tied to concrete passage content. This reshapes an educational multiple-choice resource into a subjective-answer QG dataset.

When a source item is multiple-choice, we follow the same conversion principle across datasets: the correct option becomes the intended answer and distractors are not shown to the generator. We also

Table 5: Annotated dataset split information after option cleaning.

Dataset	Source / Domain	Train	Dev	Test	Total	Contexts
EQG-RACE	Exam-style reading comprehension	12,698	1,028	944	14,670	9,783
DREAM	Dialogue-based reading comprehension	6,107	2,035	2,036	10,178	6,424
FairytaleQA	Story-based educational reading comprehension	8,548	1,025	1,007	10,580	4,261

Table 6: Major Barrett-style skill distribution in the annotated splits. Counts are shown with within-split percentages.

Dataset	Skill	Train	Dev	Test	Total
EQG-RACE	1.1 Detail Recognition	7,023 (55.3%)	649 (63.1%)	580 (61.4%)	8,252 (56.3%)
	1.4 Cause-Effect Recognition	2,999 (23.6%)	209 (20.3%)	185 (19.6%)	3,393 (23.1%)
	3.1 Inferring Supporting Details	827 (6.5%)	51 (5.0%)	61 (6.5%)	939 (6.4%)
DREAM	1.1 Detail Recognition	3,153 (51.6%)	1,111 (54.6%)	1,100 (54.0%)	5,364 (52.7%)
	3.1 Inferring Supporting Details	1,183 (19.4%)	503 (24.7%)	480 (23.6%)	2,166 (21.3%)
	1.4 Cause-Effect Recognition	648 (10.6%)	134 (6.6%)	152 (7.5%)	934 (9.2%)
FairytaleQA	1.1 Detail Recognition	4,126 (48.3%)	449 (43.8%)	465 (46.2%)	5,040 (47.6%)
	1.4 Cause-Effect Recognition	2,114 (24.7%)	281 (27.4%)	274 (27.2%)	2,669 (25.2%)
	3.5 Inferring Cause and Effect	762 (8.9%)	96 (9.4%)	75 (7.4%)	933 (8.8%)

apply additional option cleaning before evaluation. Items are excluded when the correct option only makes sense under multiple-choice logic, such as “all of the above,” “both A and B,” or options that refer to other options instead of stating an answer. Such cases are valid in multiple-choice exams but do not provide standalone answers for subjective QG.

C.3 Setup

We normalize all three datasets into the same ⟨passage, answer, target skill, question⟩ interface. Table 5 reports the evaluation split sizes used in our experiments.

D Experiments

D.1 LLM Judge Prompt

Educational metrics are computed with GPT-5 using a fixed judging prompt. For each generated question, the judge receives the passage, target answer, target Barrett skill, reference question, and generated question, and returns scores for similarity, answerability, and target-skill alignment. The prompt is shown below.

Table 7: Training hyperparameters and computational resources for **SagE-QG**.

Hyperparameter	Value
Backbone	Qwen2.5-3B-Instruct
RL algorithm	GRPO
Per-device batch size	2
Number of generations	8
Sampling temperature	1.1
Learning rate	5e-8
Max prompt length	3072
Max response length	1024
Similarity reward weight	1.0
Answerability reward weight	1.0
Skill reward weight	1.0
Interventional Grounding Gap weight	3.0
Evidence masking	Selected evidence span
Compute hardware	NVIDIA A800-SXM4-80GB GPUs
CE-RL training GPUs	2 GPUs
CE-RL step time	approx. 230–250 s / step
CE-RL compute budget	approx. 0.13–0.14 GPU-hours / step
Inference / evaluation cost	approx. 0.4–0.9 s / sample on 1 GPU

Table 8: LLM judge prompt used for educational metrics. Each generated question is evaluated independently with the same prompt.

Prompt Part	Content
System role	Act as an expert evaluator for evidence-grounded, skill-controllable reading-comprehension question generation. Return only the requested structured evaluation.
Inputs	The judge receives the reading passage, target answer, target Barrett skill, reference question, and generated question. The target-skill field includes the skill label, skill name, skill category, and skill description.
Similarity	Judge whether the generated question asks for the same underlying information as the reference question, focusing on semantic intent, required reasoning, answer target, and specificity rather than identical wording.
Answerability	Judge whether the generated question can be answered from the passage and whether the target answer is a valid answer. The passage is the only allowed evidence source.
Target Skill	Judge whether the generated question naturally requires the target Barrett skill. The prompt explicitly asks the judge not to reward superficial word overlap alone.
Score Scale	Each dimension is scored on the fixed scale $\{0.00, 0.25, 0.50, 0.75, 1.00\}$, where 1.00 indicates clear satisfaction of the criterion, 0.50 indicates partial or ambiguous satisfaction, and 0.00 indicates a mismatch.
Output	The judge returns separate reasoning and scores for similarity, answerability, and target-skill alignment, plus brief overall comments. The reported Joint score is the sum of the three aspect scores.
Retry Rule	If the returned evaluation cannot be parsed, we retry once with an instruction to return the same structured evaluation only and include the parse issues in the retry message.

D.2 Qualitative Case Studies

Table 10 shows three qualitative cases in a compact table layout. Each row groups the passage, target answer, and target skill into a single input cell, and then puts the italic reference question above the model outputs and the abbreviated **SagE-QG** trace on the right.

In the bottled-water case, the target skill is cause–effect inference. Qwen2.5-3B produces an answerable-looking question, but shifts the focus toward Isabelle Silverman’s commitment; SkillQG asks a literal question about Isabelle and misses the target answer. **SagE-QG** instead selects the evidence linking Peffers’ status-symbol argument to the public “I’ll have tap” action, yielding a question that preserves both the reference intent and the cause–effect reading skill.

The archery case tests prediction. Both base-lines turn the item into a recall question about what the old man did or instructed, which weakens the predicting-outcomes skill. **SagE-QG** keeps the question open-ended because its evidence trace centers on the young archer’s failed attack, the old man’s warning about his master, and the instruction to find that master. The generated question therefore asks for the likely next event, not a past event.

The fraud case illustrates context-based inference. Qwen2.5-3B asks about a consequence of on-line shopping, and SkillQG asks for the main idea, so neither question targets the meaning of “fraud.” **SagE-QG** selects the sentences about stolen card information, suspicious websites, and shoppers not receiving goods, then uses those clues to ask for the word meaning. Across the three cases, the common pattern is that exposed evidence selection helps **SagE-QG** keep the generated question aligned with the intended answer, skill, and human reference angle.

D.3 Masked Evidence Case Studies

Table 9 gives paired full rollouts and Counterfactual Rollouts for **SagE-QG**. In each case, we first run the model on the full passage, then mask the evidence spans selected by its own full rollout and regenerate under the same answer and target skill. The examples show that the selected evidence is not merely decorative: when it is removed, the question often shifts to a different focus or becomes less faithful to the reference angle.

The masked-pair cases make the intervention effect concrete. In the Peffers example, the full rollout grounds the question in the selected status-symbol and restaurant-response evidence, while the masked rollout changes the subject to Isabelle

Table 9: Masked evidence case studies. The Counterfactual Rollout removes the evidence selected by the full **SagE-QG** rollout and regenerates under the same answer and target skill.

Input and Selected Evidence	Full SagE-QG Rollout	Masked Re-Rollout
Case race-test-000389 Target Skill 3.5 Inferring Cause and Effect Target Answer To encourage them to set an example for others to follow. Selected Evidence <i>Bottled water's role as a status symbol needs to change ... when a waiter asks "what's your drink?" ... "Don't be afraid to say, 'I'll have tap'."</i>	Question Why does Peffers suggest saying "I'll have tap" loudly? Trace <i>Evidence scope:</i> checks Peffers' status-symbol claim and the restaurant advice ... <i>Evidence selection:</i> keeps the status-symbol sentence and the "I'll have tap" lines ... <i>Question design:</i> asks why Peffers wants the sentence said loudly, forcing the cause-effect inference. Reward 2.676	Question Why did Isabelle Silvrman stop using bottled water? Shift After Masking After the Peffers evidence is removed, the model grounds the question in other environmental and cost sentences and shifts from Peffers' public example-setting action to Isabelle's personal decision. Reward 1.564 Gap 1.112
Case race-test-000445 Target Skill 3.1 Inferring Supporting Details Target Answer Cheating. Selected Evidence <i>The Internet has led to a huge increase in credit-card fraud ... card information for sale ... suspicious cheap websites ... shoppers may never receive goods.</i>	Question What does the word "fraud" mean in the text? Trace <i>Evidence scope:</i> inspects credit-card fraud, stolen card information, and deceptive online shopping ... <i>Evidence selection:</i> keeps the sentences that define fraud through theft and non-delivery ... <i>Question design:</i> asks for contextual word meaning, not a literal consequence. Reward 2.463	Question What does the word "cheating" mean? Shift After Masking The masked rollout still finds online-crime context, but the reference word "fraud" has disappeared from the selected evidence. The question collapses into a shallower synonym-style prompt about the answer itself. Reward 2.069 Gap 0.394
Case race-test-000808 Target Skill 1.1 Detail Recognition Target Answer Their taxi has a flat tire. Selected Evidence <i>The girls decide, "let's tell him that our taxi has a flat tire," and the other girls agree.</i>	Question What excuse do the girls give to the teacher? Trace <i>Evidence scope:</i> locates the explicit excuse and the surrounding agreement ... <i>Evidence selection:</i> keeps the sentence containing "taxi has a flat tire" ... <i>Question design:</i> asks for the stated excuse, matching literal detail recognition. Reward 2.584	Question What do the four girls tell the teacher as their excuse? Shift After Masking The masked question remains answerable because nearby repeated context still contains the excuse, but it loses some of the original "being late" framing and receives a lower reward. Reward 2.146 Gap 0.438

Silverman's personal action. In the fraud example, removing the evidence that contains "fraud" and its contextual clues makes the model ask about "cheating" directly, reducing the lexical-inference quality of the question. The detail-recognition case is more robust because the answer is repeated nearby, but the reward still drops. Together, these cases show how **SagE-QG** uses selected evidence to stabilize the intended question angle.

D.4 Additional Controlled-QG Comparison

Table 11 reports additional results on DREAM. The adapted SoftSkillQG and Leite-Cardoso methods use the same split, Qwen2.5-3B-Instruct backbone, (P, A, S) input, training budget, decoding setup, and evaluation protocol as **SagE-QG**. The two larger models are evaluated as prompting references on the same test set.

D.5 Further Comparison with GPT-5

We report GPT-5 as a strong closed-model reference rather than a directly trained baseline in Table 12. The comparison is intentionally asymmetric: GPT-5 is a large foundation model trained at a much broader scale than the Qwen2.5 backbones used by **SagE-QG**, with stronger instruction fol-

lowing and likely substantial exposure to educational and question-answering data. It is therefore expected to perform strongly on judge-based educational metrics, especially those involving fluency, answerability, and human-preference alignment.

The scaled **SagE-QG** results provide a useful intermediate point. Moving from 3B to 7B narrows the gap to GPT-5 on all educational metrics, while the trained specialist remains stronger on reference-based metrics. This pattern suggests that GPT-5 benefits from broad foundation-model capability, whereas **SagE-QG** benefits from task-specific training on the target QG distribution. The comparison therefore highlights a trade-off between general closed-model capability and specialized, auditable control for educational QG.

D.6 Evidence Sensitivity Details

Table 13 extends Figure 4 by reporting the full evidence-removal Interventional Grounding Gap statistics. A larger gap and a higher positive-gap ratio indicate stronger sensitivity to the selected evidence. For the Base/Instruct row, we use a compact one-shot format prompt rather than the original zero-shot setting, because the zero-shot model does not reliably select evidence or follow the required

Table 10: Three qualitative case studies on EQG-RACE. The input cell combines the passage, target answer, and target skill; the output cell lists the italic reference question, model outputs, and an abbreviated **SagE-QG** reasoning trace.

Input	Reference Question, Model Outputs, and SagE-QG Trace
Passage <i>Downing the last drop of an expensive famous-brand H2O and remembering to throw the empty bottle in the recycling bin makes you feel pretty good about yourself. Yet even when the bottles are recycled, there are all kinds of other consequences of swallowing bottled water, says Melissa Peffers, the air-quality program manager for Environmental Defense. The containers are often filled in faraway lands, then shipped long distances. Bottled water’s role as a status symbol needs to change, Peffers points out. So when a waiter at an expensive restaurant offers, “And what’s your drink?” that’s no reason to forget your conviction. “Don’t be afraid to say, ‘I’ll have tap.’ Say it loud enough that the other tables nearby can hear you,” Peffers says.</i> Target Answer To encourage them to set an example for others to follow. Target Skill 3.5 Inferring Cause and Effect	Reference Question <i>Why does Peffers ask people to say “I’ll have tap” loudly?</i> Qwen2.5-3B According to Melissa Peffers, what is Isabelle Silverman’s commitment, and why does Peffers suggest Isabelle should encourage others to follow her example? SkillQG What does Isabelle Silverman want to do? SagE-QG Why does Peffers suggest saying “I’ll have tap” loudly? SagE-QG Trace <i>Evidence scope:</i> checks Peffers’ claim that bottled water’s status should change, then the restaurant prompt and “I’ll have tap” advice ... <i>Evidence selection:</i> keeps the status-symbol sentence plus the “what’s your drink” and “I’ll have tap” lines; drops the dessert detail ... <i>Question design:</i> ask why Peffers suggests saying it loudly, so the reader infers the intended effect: encouraging others to follow.
Passage <i>Long long ago, there was a man who was good at shooting arrows. He asked the king to declare him the greatest archer of the country. The king was about to do so when an old servant said, “Wait, sir. I know an old man who lives in the forest. He is a much greater archer. Let this young man go to him and learn from him for at least three years.” The man could not believe what the servant said, but he went. For three years he learned from the old man. One day, when he thought he had learned everything, the thought arose in him: “If I kill this old man, then I will be the greatest archer.” The old man stopped the arrow and said, “My master is still alive, and I’m nothing before him. You must be with him for at least thirty years. Go fast! Find the old man!”</i> Target Answer The young archer set off to look for the old man’s master. Target Skill 3.7 Predicting Outcomes	Reference Question <i>What do you think would most probably happen next?</i> Qwen2.5-3B What did the old man instruct the young archer to do after the young archer nearly killed him? SkillQG What did the old man do after the young man shot an arrow at him? SagE-QG What is likely to happen next? SagE-QG Trace <i>Evidence scope:</i> inspects the young archer’s thought of killing the old man, the failed attack, and the old man’s warning about a stronger master ... <i>Evidence selection:</i> keeps the intention to kill, the warning that the master is alive, and the command to “find the old man” ... <i>Question design:</i> ask for the most likely next event, prompting a prediction that the young archer will seek the master’s master.
Passage <i>The Internet has led to a huge increase in credit-card fraud. Your card information could even be for sale in an illegal web site. Web sites offering cheap goods and services should be regarded with care. On-line shoppers who enter their credit-card information may never receive the goods they thought they bought. The thieves then go shopping with your card number, or sell the information over the Internet. Computers hackers have broken down security systems, raising questions about the safety of online business.</i> Target Answer Cheating. Target Skill 3.1 Inferring Supporting Details	Reference Question <i>What is the meaning of “fraud”?</i> Qwen2.5-3B What is a potential consequence of entering your credit-card information on an online shopping site, as mentioned in the passage? SkillQG What is the main idea of the passage? SagE-QG What does the word “fraud” mean in the text? SagE-QG Trace <i>Evidence scope:</i> looks at the credit-card fraud sentence, stolen card information, suspicious cheap websites, and shoppers not receiving goods ... <i>Evidence selection:</i> keeps the fraud sentence and the clues about illegal sale, deceptive websites, and unpaid goods ... <i>Question design:</i> ask for the contextual meaning of “fraud,” requiring the reader to infer cheating from theft and deception clues.

structured output format.

D.7 Human Alignment of the LLM Judge

To validate the educational metrics, we compare LLM judge scores with ratings from three volunteer annotators on 100 sampled generated questions. The human rating protocol uses the same three dimensions as the fixed prompt: answerability, target-skill diagnosis, and human-like question design. Table 14 reports agreement after pooling rated aspect scores. The bias term is computed as $\text{mean}(\text{human score} - \text{GPT score})$ over all rated

aspects; therefore, a positive value indicates slight human leniency relative to GPT-5. The three annotators were volunteers recruited by the authors, received the same rating rubric as the LLM judge, consented to the use of their anonymous ratings for this validation, were not paid, and did not provide personal or sensitive information.

The agreement is strongest under tolerant matching: 87.3% of pooled ratings fall within one 0.25 scoring step and 96.7% fall within 0.50. The MAE is 0.170, below one scoring step. The near-zero positive bias (0.047) means that human ratings are

Table 11: Additional controlled-QG and prompted-model comparisons on DREAM.

Model	Setting	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Qwen3.6-35B-A3B	35B (3B active), prompting	17.5043	45.1225	26.1851	43.6282	0.7439
DeepSeek-R1-Distill-Qwen-32B	32B, prompting	15.0986	43.5900	22.9971	42.2504	0.7482
Leite and Cardoso (adapted)	Qwen2.5-3B, SFT	27.6013	54.6523	34.8045	53.8517	0.7836
SoftSkillQG (adapted)	Qwen2.5-3B, soft prompt + SFT	27.6306	54.7550	34.8223	53.9740	0.7862
Sage-QG	Qwen2.5-3B, SFT + CE-RL	29.7274	56.2337	36.4268	55.3622	0.7996

Table 12: Further comparison with GPT-5 on EQG-RACE. GPT-5 is reported as a strong closed-model few-shot reference, while **Sage-QG** rows are task-trained specialist models. Skill, Answer., and Human Sim. are LLM judge scores on a 0–1 scale, and Joint is their sum. Bold marks the best result and underline marks the second best.

Model	Setting	Educational Metrics				Reference-Based Metrics			
		Skill ↑	Answer. ↑	Human Sim. ↑	Joint ↑	BLEU ↑	ROUGE-1 ↑	ROUGE-L ↑	BERTScore ↑
Sage-QG	3B trained	0.7644	0.6768	0.5611	1.9724	16.0266	42.9198	41.0802	0.7542
Sage-QG	7B trained	<u>0.8151</u>	<u>0.7442</u>	<u>0.6350</u>	<u>2.1942</u>	18.2148	45.5638	43.8313	0.7702
GPT-5	few-shot	0.9447	0.9096	0.7811	2.6354	8.7345	36.2656	31.9147	0.7501

only slightly more lenient than GPT-5 on average.

Pooled concatenates ratings from the three evaluation dimensions before correlation, whereas Joint first sums the three dimensions for each question. Table 15 reports Pearson and Spearman correlations, including a cross-evaluator check with DeepSeek-V3.

E Discussion

E.1 Do Reference-Based Metrics Still Matter?

Text-similarity metrics remain useful, with a diagnostic role in our evaluation. BLEU, ROUGE, and BERTScore are inexpensive, reproducible, and sensitive to whether a generated question stays close to the reference distribution. In our reward–metric analysis, the similarity reward correlates strongly with lexical and semantic metrics, suggesting that these scores still help detect off-distribution or under-specified generations. They are therefore valuable as sanity checks for fluency, topicality, and reference consistency.

Their limitation is that educational QG is often many-to-one: the same passage, answer, and skill can support multiple valid human-like questions. A

low-overlap question may still be answerable, skill-aligned, and supported by good selected evidence, while a high-overlap question can copy surface wording without testing the intended reading skill. For this reason, reference-based metrics should be reported alongside educational metrics, Interventional Grounding Gap diagnostics, and human-alignment validation.

E.2 Closed-Source Models and Specialized Small Models

The comparison with GPT-5 raises a broader deployment question for educational QG. Large closed-source foundation models are trained at much larger scale, with stronger instruction following and likely broader exposure to educational and question-answering data. As a result, they can provide strong out-of-the-box quality, especially on educational metrics that reward fluency, answerability, and alignment with human preferences. However, this performance comes with practical constraints: inference can be costly, deployment depends on external services, model behavior may change across versions, and the generation process is harder to audit.

Table 13: Evidence sensitivity analysis. The Base/Instruct row uses a compact one-shot format prompt to elicit parseable selected evidence; the remaining rows use their trained trajectory format.

Model	Full $\uparrow$	Masked	Gap $\uparrow$	Pos. $\uparrow$	Neg. $\downarrow$	Δ Ans.	Δ Skill	Δ Sim.
Base (1-shot)	2.0323	1.9201	0.1122	0.5975	0.4025	0.0489	0.0021	0.0277
+ SFT	2.1784	2.0548	<u>0.1237</u>	<u>0.5806</u>	0.4183	0.0378	0.0065	0.0299
+ GRPO	2.1829	2.1337	0.0491	0.5511	0.4489	0.0034	0.0000	0.0162
SagE-QG	2.2260	1.9776	0.2484	0.6896	0.3104	0.1089	0.0011	0.0509

Table 14: Human-alignment validation for the LLM judge on 100 sampled generated questions.

Metric	Value
Exact agreement	0.487
Within ± 0.25	0.873
Within ± 0.50	0.967
MAE	0.170
Bias (human-GPT)	0.047

ward interpretable and avoiding excessive hyperparameter tuning. On the evaluation side, the LLM judge can be further strengthened with lightweight local multi-agent cross-verification, where several specialized judges independently assess answerability, skill alignment, and human similarity before aggregating the final score.

Specialized small models offer a complementary direction. Although **SagE-QG** does not match GPT-5 on all educational scores, it provides competitive performance with lower-cost local deployment, explicit evidence selection, and task-specific reward optimization. These properties are important in educational settings, where systems may need to generate questions repeatedly, preserve data privacy, and explain why a question was produced. The 3B-to-7B trend further suggests that the gap to closed-source foundation models can be narrowed by scaling specialist backbones, pseudo-evidence construction, and task-specific training data, while retaining the controllability and auditability of a specialized QG system.

Future Work. Small-model QG still has substantial room for improvement. Future work should scale counterfactual evidence training across all datasets and distill high-quality feedback from stronger models into smaller open models. On the training side, **SagE-QG** currently uses a fixed maximum amount of selected evidence to simplify reward design. Future work could study more compact selected-evidence sets while keeping the re-

Table 15: Correlation-based agreement on the human-validation sample.

Agreement	Pooled r	Pooled ρ	Joint r	Joint ρ
Human-Human	0.3706	0.3588	0.5109	0.5006
Human-GPT-5	0.5096	0.5054	0.6462	0.6279
GPT-5-DeepSeek-V3	—	—	0.8112	0.8424